Computers, Materials & Continua

Tech Science Press

Doi:10.32604/cmc.2025.061359

ARTICLE

Causal Representation Enhances Cross-Domain Named Entity Recognition in Large Language Models

Jiahao Wu[1,2], **Jinzhong Xu**[1], **Xiaoming Liu**[1,*], **Guan Yang**[1,3] **and Jie Liu**[4]

[1]School of Artificial Intelligence, Zhongyuan University of Technology, Zhengzhou, 450007, China
[2]School of Computer Science, Zhongyuan University of Technology, Zhengzhou, 450007, China
[3]Zhengzhou Key Laboratory of Text Processing and Image Understanding, Zhengzhou, 450007, China
[4]School of Information Science and Technology, North China University of Technology, Beijing, 100144, China

*Corresponding Author: Xiaoming Liu. Email: ming616@zut.edu.cn

Received: 22 November 2024; Accepted: 17 February 2025; Published: 16 April 2025

ABSTRACT: Large language models cross-domain named entity recognition task in the face of the scarcity of large language labeled data in a specific domain, due to the entity bias arising from the variation of entity information between different domains, which makes large language models prone to spurious correlations problems when dealing with specific domains and entities. In order to solve this problem, this paper proposes a cross-domain named entity recognition method based on causal graph structure enhancement, which captures the cross-domain invariant causal structural representations between feature representations of text sequences and annotation sequences by establishing a causal learning and intervention module, so as to improve the utilization of causal structural features by the large language models in the target domains, and thus effectively alleviate the false entity bias triggered by the false relevance problem; meanwhile, through the semantic feature fusion module, the semantic information of the source and target domains is effectively combined. The results show an improvement of 2.47% and 4.12% in the political and medical domains, respectively, compared with the benchmark model, and an excellent performance in small-sample scenarios, which proves the effectiveness of causal graph structural enhancement in improving the accuracy of cross-domain entity recognition and reducing false correlations.

KEYWORDS: Large language model; entity bias; causal graph structure

1 Introduction

In recent years, LLMs such as GPT-3 [1] have demonstrated remarkable performance in zero-shot and few-shot tasks, advancing the research on CD-NER [2]. However, CD-NER still faces the challenge of scarcity of large scale annotated data in specific domains. This scarcity limits the model's ability to learn features relevant to the target domain, leading to reduced accuracy in entity recognition. Additionally, it exacerbates issues related to causal and spurious correlations within NER tasks, which traditional methods struggle to effectively address. The emergence of LLMs has significantly alleviated this dependency. Although LLMs have exhibited excellent generalization capabilities in NLP tasks through contextual learning [1] and chain-of-thought [3] techniques, their performance in cross-domain NER tasks still requires improvement. Current research on cross-domain tasks primarily models the interaction between domains using two methods: domain adaptation based on pre-trained language models and prompt-based approaches leveraging LLMS. Domain adaptation methods enhance target domain performance by sharing source domain knowledge and domain-invariant features, but their generalization ability relies on the similarity between the source

and target domains. When there is a significant disparity, model performance tends to decline markedly. Prompt-based methods, on the other hand, improve target domain performance by designing domain-specific prompts to activate the latent capabilities of LLMs [4]. However, their effectiveness depends on the choice and design of prompts, and they have limited generalization capacity to new domains.

Furthermore, selective biases induced by label and contextual correlations across domains can lead to model overfitting to non-representative features, ignoring the true causal relationships and thereby constraining the model's generalization capability and prediction accuracy. As illustrated in Fig. 1, "John Preskill" should be recognized as a "scientist" rather than merely a "Person", and the reason for this problem is that LLMs will incorrectly learn the correlation between "John Preskill" and the "person" during the pre-training process, but this correlation is a spurious correlation problem. Therefore, in complex cross-domain tasks, causal and spurious correlations are critical issues [5]. Causal correlations are stable and interpretable, whereas spurious correlations fluctuate with changes in the environment. Recent studies have attempted to mitigate the impact of spurious correlations using causal inference methods [6–8], optimizing the model's predictive capabilities, reducing the negative effects of confounding factors, and enhancing performance in cross-domain tasks.

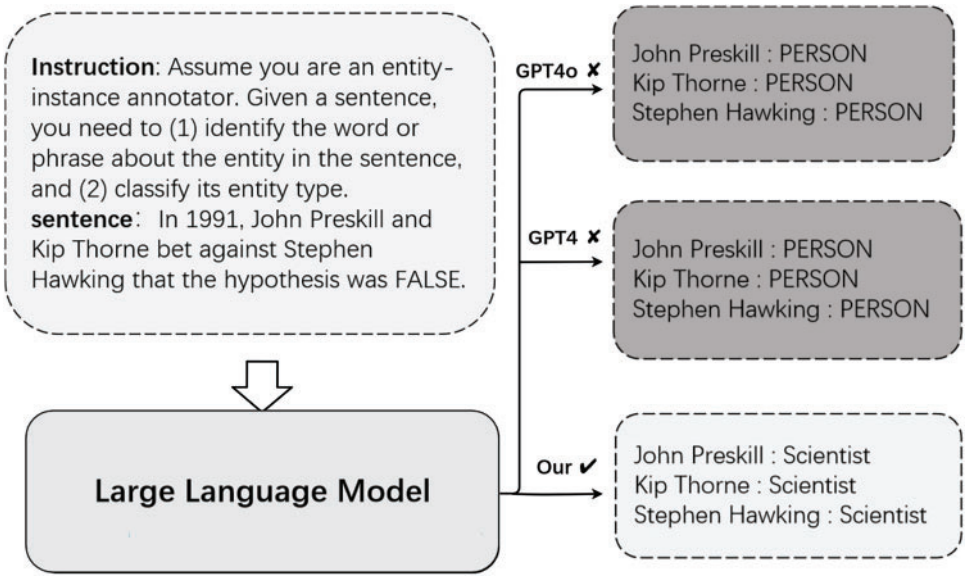

Figure 1: Causal learning helps models alleviate entity bias

Despite the effectiveness of these methods in reducing the over-prediction issues in cross-domain tasks, they still face several challenges. Firstly, there is a lack of thorough exploration of causal relationship consistency across domains. Traditional approaches are often applied to model training in specific scenarios and fail to fully leverage underlying structured knowledge [9]. Secondly, the issue of entity bias [10] is difficult to effectively filter out. The variation in entity information across different domains can affect LLMs, leading them to rely on biased parameters and make unreliable predictions [11]. Additionally, LLMs parameters are not accessible and their logical structure is uncalibrated, which hampers the capture of causal relationships in cross-domain tasks, resulting in inadequate adaptability to diverse domain data.

To address the aforementioned challenges, this paper proposes a CD-NER method based on the enhancement of causal graph structure. The model is designed to meet the different requirements of different domains. The model generates fused semantic features by migrating the original semantic information and

domain invariant feature knowledge from the source domain model to the target domain and combining them with the target domain. In addition, the fusion semantic features are then used to build a causal graph with the target domain labels to capture the causal semantic features between the fusion semantic features and the target domain labels, so as to explain the causal relationship between them. After that causal intervention and counterfactual strategies are used to generate cross-domain invariant causal structure predictions, which enhances the ability of LLM to recognize causal relationships and improves its performance in cross-domain entity information detection. The main contributions are as follows:

1. This paper proposes a method that fuses multiple domains to identify and leverage causal relationships, enhancing large language model performance by mitigating semantic interference, causal inconsistency, and entity bias in cross-domain tasks.
2. Experimental results on several domain-specific NER datasets validate the effectiveness of this approach.

2 Related Works

This section will introduce related works in the areas of cross-domain NER, large language models, and causal invariant learning.

2.1 Cross-Domain Named Entity Recognition

Current CD-NER models primarily employ two approaches: domain adaptation and contrastive learning. Mou et al. [12] introduced the basic concepts and ideas of domain transfer, leveraging shared knowledge from the source domain to enhance the performance in the target domain. However, Tang et al. [13] pointed out that previous methods neglected domain-specific information, leading to conflicts in entity types. To address this, they built upon the work of Li et al. [14] by employing entity discrimination tasks and entity-aware training settings to mitigate the negative impacts of domain-specific entity type dependencies. Wang et al. [15] explored multi-domain adaptation by setting up linear and layers for each domain. Nevertheless, with an increase in the number of domains, the model tends to become more complex and difficult to train.

In the context of contrastive learning methods, Das et al. [16] suggested using Gaussian embeddings and contrastive learning to improve the accuracy of few-shot NER. However, this approach overlooked the integrity of entities. Xu et al. [17] proposed guided momentum contrastive learning, which improves the accuracy of cross-domain NER by guiding the learning process through EB and LB.

In comparison, previous methods have failed to adequately leverage the causal relationship between features and labels to optimize predictive performance. This shortfall has resulted in suboptimal handling of entity bias issues. Moreover, these methods have not fully explored the rich knowledge embedded in LLMs.

2.2 Large Language Models

Recent research has increasingly explored the use of LLMs for IE tasks [1]. LLMs have demonstrated excellent performance across a wide range of NLP tasks, particularly in zero-shot environments. By leveraging instruction tuning, they have significantly enhanced performance, showcasing their great potential in IE tasks [1]. However, recent studies indicate that LLMs still face challenges when dealing with complex cross-domain tasks [18]. Qin et al. [19] found that ChatGPT shows limited performance in zero-shot sequence labeling tasks. Fei et al. [20] suggested that the choice and order of examples in context learning may lead to prediction biases. Wang et al. [21] argued that entity bias significantly affects large models, causing these LLMs to rely on biased parameters, resulting in unreliable predictions. Ye et al. [22] proposed using data augmentation techniques to improve LLMs' capabilities in few-shot NER by enhancing original data with

context and entity-level augmentations to utilize the unique characteristics of NER tasks. Bernal et al. [23] pointed out that LLMs perform worse than traditional pre-trained language models in few-shot biomedical relation extraction tasks. Zhang et al. [24] observed that even with instruction tuning, LLMs' performance in certain IE tasks still struggles to surpass that of pre-trained language models.

2.3 Causal Invariant Learning

Causal invariant learning is a common approach used to address domain adaptation and domain generalization problems in cross-domain transfer learning. Domain generalization is crucial for learning domain-invariant causal knowledge. Tang et al. [25] proposed a scene graph generation framework based on causal reasoning. By constructing causal graphs, they extracted counterfactual causal relationships to eliminate the impact of biases. However, this method did not consider the biases implied by cross-cultural differences. Lin et al. [26] categorized biases into intra-domain and cross-domain types. They proposed constructing causal models using a hierarchical Bayesian deep model to calculate causal effects and eliminate both intra-domain and cross-domain biases through causal intervention. To address selection bias and distributional bias in data, Ren et al. [8] developed a framework for covariance and variance optimization to learn the causal relationships between features and targets, minimizing covariance to obtain causal effects and resolve data bias issues between NER and relation extraction. Yang et al. [10] replaced contextual information with causal intervention to uncover the primary causal relationships in data from a causal reasoning perspective. Cao et al. [27] suggested using structural causal models as analytical tools to identify hidden potential risks in exploratory tasks, thus reducing data biases.

3 Problem Definition

Assuming an input sequence $X = \{x_1, x_2, \ldots, x_n\}$, where n is the sequence length and each x_i represents a token in the input sequence, the corresponding output sequence is $Y = \{y_1, y_2, \ldots, y_n\}$, where each y_i is the predicted outcome for x_i. Each predicted outcome is defined over a real category set, and it is selected from the real category set of the input sequence X. In this paper, a domain D generally consists of two parts: a feature space X and a marginal probability distribution $P(X)$, if two domains exist, they may have different feature spaces or different marginal probability distributions. This paper considers a source domain D_S, a target domain D_T. The method defines the source domain dataset as $D_S = \{(x_{s_1}, y_{s_1}), (x_{s_2}, y_{s_2}), \ldots, (x_{s_n}, y_{s_n})\}$ where $x_{s_i} \in X_S$ is an instance and y_{s_i} is its label. For CD-NER task, using the data from the source domain D_S, can train a predictive model backbone. The goal is to acquire knowledge that can be generalized across multiple domains and apply it to the target domain data set D_T, ensure that the model, given an input sequence, can maximize the conditional probability distribution of the label sequence:

$$P(y \mid x; \theta) = \prod_{t=1}^{T} P(y_t \mid x_t, y_{(1:t-1)}; \theta) \tag{1}$$

4 Model

CD-LM based on causal graph structures. The method constructs causal graphs to capture cross-domain invariant causal representations between the feature representations of text sequences and the labeled sequences, thereby improving the performance of LLMS in target domain entity recognition. As shown in Fig. 2, the model first introduces a feature fusion module to obtain label structure information between the source domain and the target domain. By employing GCN alignment techniques, the model acquires cross-domain invariant structural information. Subsequently, the model utilizes causal learning to capture potential causal relationship features within text sequences and enhances them into cross-domain

invariant causal graph representations. Through counterfactual strategies, the model proactively adjusts and evaluates the causal effects among features, eliminating the confounding factors caused by entity bias, and ensuring the reliability and consistency of causal relationships. This approach allows the model to accurately capture the causal structures within text sequences. Finally, using prompts and LLMs, the model enhances its ability to represent cross-domain invariant features, thereby improving the learning and prediction of causal relationships in text. This enhancement boosts the detection accuracy of entity information across different domains.

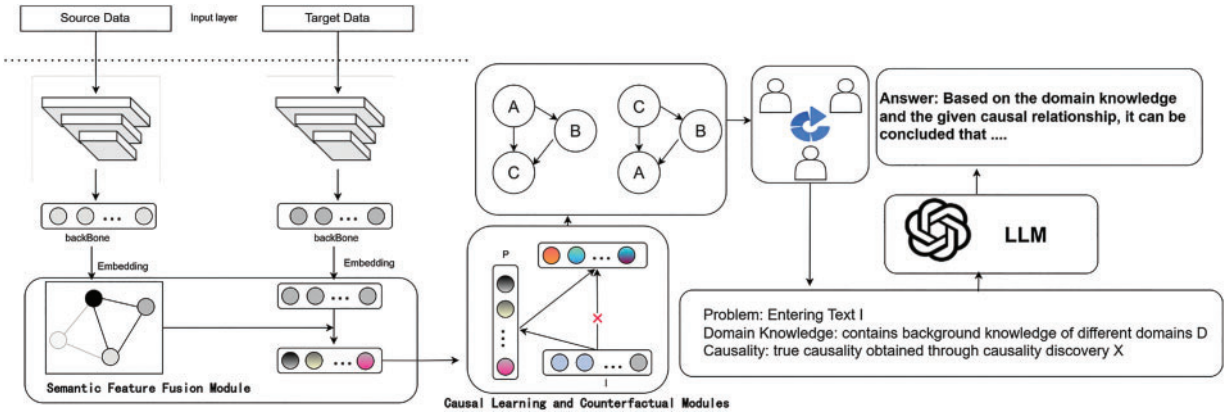

Figure 2: Overall framework of CDLM

4.1 Semantic Feature Fusion Module

Due to potential mismatches between the labels in the source domain and the target domain datasets, models trained on the source domain cannot be directly applied to the target domain. The differences in label distribution necessitate adjustments and adaptations in the target domain. The method need to calculate the probability distribution of mapping the target domain labels to the source domain labels $p(y_s \mid y_t = y)$. In this study, the method refer to the method by Zheng et al. [28], where the predictions of the source model for all samples of entity types in the target domain are averaged. The specific calculation formula is as follows:

$$p(y_s \mid y_t = y) = |\mathscr{D}_T^y|^{-1} \sum_{(x_t, y_t) \in \mathscr{D}_T^y} \text{softmax}\left(f_s(x_t)\right) \tag{2}$$

Here, y_s and y_t represent the labels in the source domain and the target domain, respectively, f_s represents the pre-trained language model back given the source domain dataset D_s, corresponds to backbone in Fig. 2, $|\mathscr{D}_T^y|$ represents the number of training samples with true labels in the target domain. Subsequently, a source domain graph $G_S(V_S, E_S)$ is constructed, where the graph nodes V_s represent entity labels and the edges, represent semantic similarity. Entity labels with similar semantic features will have similar probability distributions. Based on this characteristic, the proposed node representation of in the graph is given by:

$$\bar{t}_s^y = \left[p\left(y_s^{(1)} \mid y_t = y\right), \ldots, p\left(y_s^{(i)} \mid y_t = y\right)\right] \tag{3}$$

$\bar{t}_s^y \in \mathbb{R}^{|y_s|}$ represent the original semantic feature representation of each graph node, i.e., the node's original, unprocessed or transformed feature representation in the graph, $|y_s|$ is the number of entity labels in the source domain. It is necessary to normalize the original semantic feature representation to obtain

more accurate data. This paper adopts the Wasserstein distance [28] as the distance function to calculate the distance between nodes. Therefore, the edges can be represented as:

$$t_s^y = \frac{\overline{t}_s^y \cdot |y_t|^2}{\sum_{y_1,y_2} W_p\left(\overline{t}_s^{y_1}, \overline{t}_s^{y_2}\right)} \tag{4}$$

where, t_s^y represents the linguistic feature representation of each node after normalization. The function is used to calculate the Wasserstein distance between any two nodes y_1 and y_2. In the process of graph construction, an edge is added between two nodes only if the distance $W_p\left(\overline{t}_s^{y_1}, \overline{t}_s^{y_2}\right)$ is smaller than a threshold value δ. After constructing the source domain graph, in order to further enhance the relationships between labels, a mechanism based on label-guided attention is employed to strengthen the effect of deterministic labeling. Specifically, given a sentence X in the target domain with the actual label sequence, where $X = \{x_1, x_2, \ldots, x_n\}$, each word vector $\mathbf{w}_i \in \mathbb{R}^d$ can be generated from the previous model. Here d represents the dimensionality. By employing this mechanism, the nodes in the source domain graph $\overline{t}_s^y$ are replaced with probability distributions based on the label l_i. The detailed formula is as follows:

$$q_j = h_j W_p + b_p \tag{5}$$

$$l_i = \sum_j \frac{\exp\left(q_j C_i^T\right)\left(w_i W_p + b_p\right)}{\sum_j \exp\left(q_j C_i^T\right)} \tag{6}$$

where W_p and b_p represent the weight and bias, respectively, w_i represents the original embedding of the j-th token in the sentence, and c_i represents the randomly initialized label embedding. After constructing the label-based source domain graph, a semantic and label fusion method is employed. This method integrates the learned graph structure embedding into the context of each word in the sentence. Subsequently, a Graph Convolutional Network is used for message passing to aggregate the semantic and similarity features between nodes. This process enhances the actual embedding representation of labels. The detailed formula is as follows:

$$\overline{l} = \mathrm{GCN}(l) \tag{7}$$

where l represents the aggregated node representation of specific labels. This aggregated representation of specific labels is integrated into the context of the target domain to blend the relevant knowledge features from both the source and target domains. This process facilitates the learning of semantic similarity features from the source domain and combines them with the features of the target domain. Consequently, it results in an enhanced contextual semantic representation in the target domain. The detailed formula is as follows:

$$\overline{w}_i = w_i + \left(\frac{\sum_i \exp\left(q_j^T \overline{l}_i\right)\overline{l}_i}{\sum_k \exp\left(q_j^T \overline{l}_i\right)}\right)\overline{W}_p + \overline{b}_p \tag{8}$$

where $\overline{w}_i$ represents the token embedding of the i-th token after label fusion. $\overline{W}_p$ and $\overline{b}_p$ denote the corresponding projection weight and bias, respectively. To ensure that the semantic features of the fusion model are more accurately focused on the correct entity types, a BCE loss function is employed. The specific loss calculation is as follows:

$$L_{\text{fusion}} = \mathrm{BCE}\left(\mathrm{Linear}\left(\left[\overline{w}_i, \ldots, \overline{w}_i\right]\right), Y\right) \tag{9}$$

where [;] represents the concatenation operation, Y represents the ground truth label of the sentence. Through the loss calculation this approach not only accurately learns the relevant entity information from both the source and target domains but also effectively integrates the entity information from both domains.

4.2 Causal Learning and Counterfactual Modules

This paper constructs a structural causal model to describe the causal learning module within the CDLM. In NER tasks, a causal graph can be used to represent the causal relationships among different variables. This graph is represented by a Directed Acyclic Graph, where node I represents the fused semantic features generated by the text sequence in the target domain, which is composed by fusing together the semantic features output from the target domain and the graph structure constructed in the source domain. Node E represents focusing only on entity semantic features in the target domain, node C represents focusing only on confounding factors in the target domain, node P represents representing prior knowledge in the target domain, node D represents domain information in the target domain, and node Y represents the output of labeling. Specifically, the causal graph is illustrated in Fig. 3: where Fig. 3a represents shows an example of entity bias, Fig. 3b depicts the causal relationship between the variables and Fig. 3c shows a counterfactual operation. Specifically, (a) illustrates an example of entity bias. The name "Ludwig van Beethoven" is easily associated with a "concert hall", but not all musicians are related to a "concert hall", and not every instance of "Ludwig van Beethoven" should be identified as a Musician; it could also be classified as a Person. In the causal graph illustrated in (b), domain information D (such as the entity type musician rather than person), prior knowledge P represents the fused semantic features from the previous step, entity information E (such as Ludwig van Beethoven), predicted label Y (such as the type musician), and confounding factor C represents the confounding factor introduced by entity bias. Specifically, as shown in Fig. 3a, the name "Ludwig van Beethoven" can be easily associated with "concert hall", but not all musicians are associated with "concert hall" is associated with "concert hall" and not every instance of "Ludwig van Beethoven" should be recognized as a musician, and therein lies the influence of confounding factors.

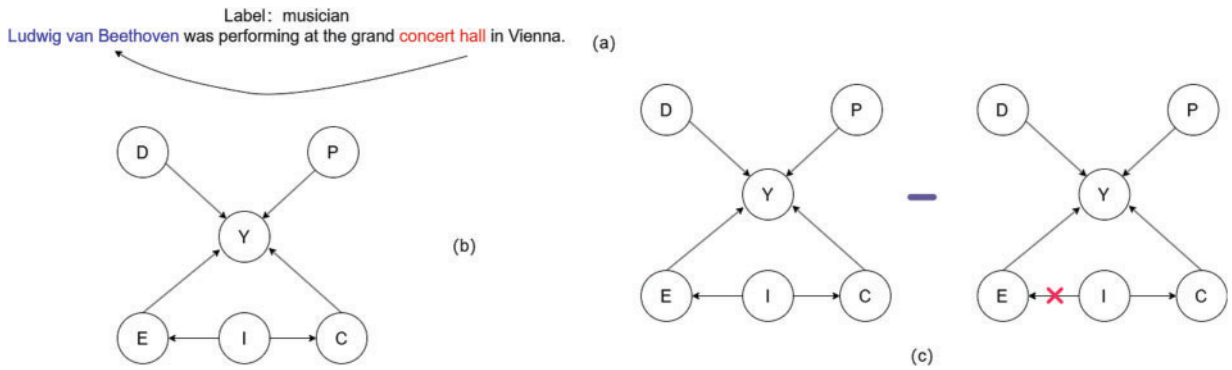

Figure 3: Causal model diagram

Since the input text I consists of entity information E and confounding factors C, it includes causal paths $I \rightarrow E$ and $I \rightarrow C$. In the generation process, the feature Y is directly influenced by entity information E, confounding factors C, prior knowledge P, and domain knowledge D. Therefore, it can be represented by the causal paths $P \rightarrow Y$, $E \rightarrow Y$, $C \rightarrow Y$ and $D \rightarrow Y$. The paths $P \rightarrow Y$, $C \rightarrow Y$ and $D \rightarrow Y$ may affect the generation results and introduce issues such as linguistic bias, irrelevant bias, and entity bias. Based on the causal graph, it is possible to assess the causal relationships between the context and the entity. Setting the

input feature as i, entity information as e, confounding factors as c, prior knowledge as p, and domain feature as d, the expression for the generated feature Y can be represented as:

$$Y_{p,e,c,d} = P(Y \mid do(P = p), do(E = e), do(C = c), do(D = d)) \tag{10}$$

where the do-operator is the back-door criterion. To more accurately calculate the direct causal effect of entity information E on the output Y, a method of controlling variables is employed. First, the total direct causal effect after counterfactual reasoning is calculated. This is represented using the English letters directly:

$$E_{\text{total}} = Y_{e,c,p,d} - Y_{e^*,c^*,p^*,d^*} \tag{11}$$

Here, e^*, c^*, p^*, d^* are represented in the counterfactual scenario to simulate the baseline state of each variable in the absence of intervention. To further obtain the direct causal effect of E, excluding the influence of prior knowledge, confounding factors and domain feature, the calculation can be expressed using the following formula:

$$E_{rpd} = Y_{e^*,c,p,d} - Y_{e^*,c^*,p^*,d^*} \tag{12}$$

By controlling the variable $E = e^*$, the potential causal relationships of E can be completely excluded. By comparing the baseline states and non-baseline states of c, p, d and c^*, p^*, d^* for the baseline states, the direct impact of these three variables on the output Y can be directly estimated. Therefore, the direct causal effect of the entity feature E on Y can be obtained as follows:

$$E_{\text{total}} - E_{rpd} = Y_{e,c,p,d} - Y_{e^*,c,p,d} \tag{13}$$

This difference represents the effect of entity feature E on Y when R, P and D are held constant, considering only the impact of E on Y. This allows for the direct calculation of the causal effect of E on Y, excluding the potential confounding effects introduced by r, p, d.

As shown in Fig. 3c, a counterfactual operation is demonstrated, which directly calculates the direct causal effect of entity feature E on y. Specifically, in the counterfactual process, by cutting off the edges, and thus the connection between different nodes, the node E is re-assigned a value so that the value of node E no longer relies on the influence of the parent node I, and this intervening operation can be called a counterfactual. Counterfactual is to reflect "what is the difference in the result for different variables". And the idea of counterfactual can guide the model to think: what is the key information that determines the output of the entity in the target domain? Specifically, a counterfactual represents a modification of a value such as X to speculate on the possible outcomes of Y. The formula is as follows:

$$P(Y = y \mid do(X = x')) = \sum_z P(Y = y \mid X = x', Z = z) \cdot P(Z = z) \tag{14}$$

where $do(X = x')$ means that we modify the value of X to x', and Z denotes the external factors affecting X and Y, such as confounders and domain characteristics. By constructing hypothetical scenarios, counterfactual analysis can assess the impact of input variables on the output results, thereby enhancing the interpretability and reliability of the model, especially in the context of feature transfer and complex interactions. This paper employs both explicit and implicit counterfactual strategies.

Explicit Counterfactuals: Explicit counterfactuals involve direct intervention or control operations to transform specific variables into counterfactual scenarios, thereby evaluating the direct effect of particular features on dependent variables. To clearly ascertain the impact of specific entity information, a masking

operation is performed on feature vectors. By observing the changes in the model's prediction results after masking out real entity information, the effect can be evaluated. The masking operation involves creating a mask vector M, and then using the Hadamard Product (element-wise multiplication) to obtain the feature vector w_i^* after masking. The specific formula is as follows:

$$w_i^* = \overline{w}_i \odot M \tag{15}$$

This method focuses on masking entity information to more clearly observe changes in the model output in the absence of such information. The specific processing workflow can be represented by the following calculation formula:

$$P(y_j \mid w_i) = P(y_j \mid w_i^*) - \alpha \cdot P(y_j \mid \overline{w}_i) \tag{16}$$

where i represents the text representation after being masked, w_i^* and w_i represent the probability representation of the j-th token. The bias used to represent the distance between the content of the generated real entity information and the original text.

Implicit Counterfactuals: Implicit counterfactuals utilize latent variable models or generative models to simulate hypothetical scenarios and capture potential causal relationships. Specifically, in each understanding module, the cross-attention mechanism dynamically divides the input into two disjoint parts (i.e., real entity information u and other information including r, c and d). The decoding module then processes these parts separately for counterfactual training. In the decoding module, the probability of generating token y_i can be represented as $P(y_j \mid e)$ and $P(y_j \mid c)$.

To ensure that the generative model reduces dependence on important labels, the method introduce a consistency loss:

$$L_e = -\sum_{i=1}^{|y|} \log\left(1 - P(y_i \mid e)\right) \tag{17}$$

where y represents the predicted result of the entity, is used to increase the probability of generating labels when focusing on unrelated tags:

$$L_{crd} = -\sum_{i=1}^{|y|} \left[\log P(y_i \mid c) + \log P(y_i \mid r) + \log P(y_i \mid d)\right] \tag{18}$$

4.3 Prompt Module

This section proposes a causality extraction module for LLMs consisting of three components: causality discovery, voting discussion iteration, and counterfactual reasoning. The following is the specific design:

Given a set of causal variables $X = \{x_1, x_2, \cdots, x_n\}$ and an outcome set $Y = \{y_1, y_2, \cdots, y_n\}$ containing explanations of the causal variables, LLMs play the role of multiple domain experts to perform causality discovery based on this variable set and outcome set, identifying causal statements denoted as $\mathbb{S} = \{\langle x_i \rightarrow y_i \rangle\}$, where $\langle x_i \rightarrow y_i \rangle$ denotes that x_i leads to y_i. The part of the prompt is: Please identify all the words or phrases in this text that may be causally related. You do not need to analyze or make judgments about the specific causal relationships involved in these words or phrases, but simply focus on extracting expressions that may indicate or imply a causal relationship. These might include words or phrases describing cause and effect, influence, result, action, correlation, etc.

Given a text I and a set of dependent variables X, LLMs based on a full understanding of the set of dependent variables X, combine information from multiple domains D to analyze and find the direct

causal relationship between the set of dependent variables X, without the need to look for indirect causality, which can be directly inferred from the direct. The part of the prompt is: Read carefully for possible causal relationships between this text and the one I've provided, extracting words or phrases with direct causal relationships based on your expertise and making sure that each causal chain accurately represents a direct causal relationship between the two variables, rather than an indirect association.

$$S = LLM(X, I, D) \tag{19}$$

where $S = \{\langle x_i \to y_j \rangle, \cdots, \langle x_i \to y_j \rangle \mid x_i, y_j \in X, Y\}$. After extracting the causal statements in LLMs, multiple sets of causal pairs are obtained $S' = \langle x_i \to x_j \rangle$, based on causal pairs, the constraint T can be specified as follows:

$$T = \{x_i \to x_j \mid (x_i, x_j) \in S'\} \tag{20}$$

With the multiple constraints M derived from the current causality discovery phase, multiple rounds of expert voting are conducted to go through to verify the reasonableness of the direct causality, and in case of conflict, multiple rounds of expert discussion are conducted, where the experts will vote individually on the issue and give specific opinions until all the experts' opinions are agreed upon. The part of the prompt is: Read this text carefully and score the results, evaluating the results in terms of accuracy, relevance, and domain adaptation, each on a scale ranging from 1 to 5 (1 being the worst and 5 being the best), with specific explanations. Specifically, the output of the LLMs is combined with the scoring-based approach in each round of discussion:

$$y' = \arg\max_{y \in Y} P(y \mid x, M) \tag{21}$$

Finally, model integrates background knowledge with the final causal structure and uses counterfactual reasoning to validate its response, avoiding hallucination. Counterfactual reasoning involves generating an alternative answer, comparing it to the original, and checking for contradictions. If inconsistencies are found, the model evaluates the alternative and refines its response; otherwise, it outputs the original answer.

4.4 Objective Optimization and Optimization Algorithms

Finally, the overall loss of the model can be expressed as:

$$L = L_{\text{fusion}} + \lambda_1 * L_e + \lambda_2 * L_{\text{crd}} \tag{22}$$

where L_{fusion} represents the binary cross-entropy loss during the feature fusion process, and L_{crd} are the losses for the causal relationship extraction task and the general relationship extraction task, respectively. By adjusting the weight coefficients λ_1 and λ_2, the influence of these two loss items on the total loss of the model can be balanced. Specifically, λ_1 corresponds to the weight for L_e, and λ_2 corresponds to the weight for L_{crd}. This approach not only balances the contributions of the two loss components but also enables the model to effectively focus on extracting causal relationship information, thereby improving overall performance.

5 Experiments

To demonstrate the effectiveness of the method proposed in this paper, the method conducted tests on five English datasets as well as a dedicated cross-domain dataset. The experimental results were analyzed from five different perspectives, including ablation studies, parameter testing, case studies, LLMs model analysis, and further extended experiments.

5.1 Datasets

This study utilizes six publicly available datasets, including CoNLL-2003 [29], BioNLP13PC [30], BioNLP13CG [30], MIT-Restaurant [31], MIT-Movie [31], and Cross-NER [32]. The CoNLL-2003 English dataset is derived from the Reuters corpus. The BioNLP13PC dataset originates from the BioNLP 2013 Shared Task, focusing on extracting pathway information from biomedical literature, while the BioNLP13CG dataset is dedicated to extracting information related to cancer genetics. The MIT-Movie dataset serves as a benchmark for enhancing text processing capabilities in the movie domain, and the MIT-Restaurant dataset is a training and testing corpus for semantic labeling in the restaurant domain. The Cross-NER dataset includes five distinct domains (politics, natural science, music, literature, and artificial intelligence), each featuring unique entity types. For ease of understanding, in the following we replace the dataset with a more concise form, shortening BioNLP13PC to PC, BioNLP13CG to CG, MIT-Restaurant to Res, and MIT-Movie to Mov. In addition, the different domains of Cross-NER will be abbreviated differently: the politics domain will be abbreviated as Pol, natural science as Sci, music as Mus, literature as Lit, and artificial intelligence as AI.

5.2 Experimental Settings

For this study, the adopted approach is based on the pre-trained language modeling framework on BERT [33]. The adopted LLM is based on LLaMA3.1-8B and GPT-3.5-Turbo-0125. After several iterations of parameter tuning, the following optimal experimental parameters are selected: stochastic gradient descent is chosen as the optimizer, the learning rate is set to 0.0001. The batch size is set to 8, and the hidden layer size is set to 768. To prevent overfitting, the dropout rate is set to 0.5. The evaluation metrics used are consistent with those used in previous studies, and the micro-F1-score is adopted as the main evaluation metric. This metric combined precision and recall across all categories to provide a more comprehensive assessment of model performance. The final result is the average of five independent runs to ensure robustness and reliability. Given the diverse characteristics of these datasets, this study designs the experiments to validate the proposed method's effectiveness in two main parts. In the first part, the CoNLL-03 dataset is used as the baseline dataset for CD-NER, with experiments conducted on the PC, CG, and Cross-NER target datasets. The second part of the experiment aims to explore the few-shot cross-domain transfer capabilities based on the CoNLL-03 dataset, selecting Cross-NER, Res and Mov as the target datasets to test the transferability.

5.3 Baselines

To validate the effectiveness of the proposed model, comparative experiments were conducted against related models on different datasets. Coach [34]: Liu et al. first detect whether tokens are slot entities to learn general patterns and then classify the slot entities, which improves the prediction accuracy in specific domains. LANER [9]: Hu et al. enhance the relationship between labels and tokens through multi-task learning, improving the transferability of label information and facilitating mutual promotion of NER tasks between source and target domains. NNShot [35]: Yang et al. train an NER model on the source domain as a feature extractor and then classify features using nearest neighbors. StructShot [35]: Building on NNShot, this method introduces structured information to enhance the model's recognition capability. LST-NER [28]: Zhang et al. model label relationships as probability distributions, constructing a label graph for cross-domain NER tasks in scenarios where the label sets of the source and target domains differ. LightNER [36]: Chen et al. improve the overall performance of NER tasks in resource-constrained environments by incorporating prompts during model training. TemplateNER [37]: Ma et al. transform NER into a large language model task through template-free prompt tuning techniques, enhancing NER performance in few-shot scenarios. CP-NER [38]: Chen et al. propose using frozen pre-trained language model parameters

and cross-domain prompt techniques to integrate knowledge from multiple domains, enhancing NER performance in the target domain and preventing performance degradation due to insufficient data from a single domain.

5.4 Results

The experimental results on various commonly used cross-domain datasets are presented in Tables 1–3, where CDLM(LLaMa) in the table represents LLM using the LLaMA3.1-8B model and CDLM(GPT) represents LLM using the GPT-3.5-Turbo-0125 model. The bold font indicates the top performance in the comparative experiments, the italicized font_indicates the second-best results, and a dash—indicates the absence of experimental results. Overall, the proposed CDLM outperforms the baseline models in both resource-rich and resource-scarce domains. Compared to state-of-the-art models, it also demonstrates significant improvements across multiple datasets and domains. The method adopted a model architecture similar to LST-NER as the foundational framework. Although LST-NER is based on a single-task framework and its performance is inferior to the multi-task LANER in several aspects, as well as showing a significant gap when compared to prompt-based models like LightNER and CP-NER, the modified CDLM demonstrates significant improvements across various dimensions.

Table 1: Experimental results of the CrossNER dataset (%)

Method	Pol	Sci	Mus	Lit	AI	PC	CG
Coach	61.50	52.09	51.66	48.35	45.15	—	—
LST-NER	70.44	66.83	72.08	67.12	60.32	87.12	82.48
LANER	71.65	69.29	73.07	67.98	61.72	—	—
LightNER	72.78	66.74	72.28	65.17	35.82	—	—
CP-NER	73.41	74.65	78.08	70.84	64.53	—	—
CDLM(LLaMa)	74.12	73.41	79.32	71.23	64.44	88.91	84.55
CDLM(GPT)	**74.37**	**73.62**	**80.01**	**71.25**	**64.90**	**89.33**	**85.32**

Note: Bold values indicate the best performance, while underlined values indicate the second-best performance.

Table 2: Experimental results for few-shot scenarios (K = 20) (%)

Method	Pol	Sci	Mus	Lit	AI	Mov	Res
TF-NER	60.15	61.22	65.38	61.26	45.23	40.12	38.53
NNShot	60.93	60.67	64.21	61.64	54.27	—	—
StructShot	63.31	62.95	67.27	63.48	55.16	—	—
TemplateNER	63.39	62.64	62.00	61.84	56.34	—	—
LST-NER	64.06	64.03	68.83	64.94	57.78	57.83	58.26
CDLM(LLaMa)	65.42	66.78	70.86	66.12	59.12	60.75	63.12
CDLM(GPT)	**65.71**	**67.08**	**71.39**	**67.02**	**59.27**	**61.06**	**63.74**

Note: This table shows experimental results for few-shot scenarios with K = 20. Bold values indicate the best performance, while underlined values indicate the second-best performance.

Table 3: Experimental results for few-shot scenarios (K = 50) (%)

Method	Pol	Sci	Mus	Lit	AI	Mov	Res
TF-NER	58.42	64.55	62.45	62.31	48.58	41.28	40.55
NNShot	66.33	63.78	67.94	63.19	59.17	—	—
StructShot	67.16	64.52	70.21	65.33	59.73	—	—
TemplateNER	65.23	62.84	64.57	64.49	56.58	—	—
LST-NER	68.51	66.48	72.04	66.73	60.69	61.25	63.58
CDLM(LLaMa)	<u>69.32</u>	<u>68.53</u>	<u>74.53</u>	<u>68.21</u>	<u>61.35</u>	<u>64.03</u>	<u>65.71</u>
CDLM(GPT)	**69.65**	**69.12**	**75.09**	**68.63**	**61.59**	**64.41**	**66.07**

Note: The table shows the experimental results for few-shot scenarios with K = 50. Bold values represent the best performance, while underlined values indicate the second-best performance.

Notably, when compared to the multi-task LANER, the CDLM exhibits substantial enhancements across the five domains of the CrossNER dataset. For instance, it achieves an increase of 2.47% in the Pol domain and 4.12% in the Sci domain, which effectively validates the efficacy of the proposed method. This method leverages the pre-trained language model's capability to effectively identify causal relationships between entities and contexts. It also utilizes the extensive corpora within the LLM to provide contextual support and deepen understanding, addressing potential limitations of LLMs in comprehending and generating content based on causal relationships. Furthermore, the proposed method does not require additional training for the LLM, thereby fully harnessing the LLM's inherent capabilities. It is adaptable to various types of LLMs and can optimize their performance across different datasets without altering the model architecture or parameters. In terms of training time and resource consumption, the method significantly underperforms CP-NER, yet achieves comparable results and exhibits notable improvements in certain domains, such as a 0.71% increase in the Pol domain and a 0.68% increase in the Mus domain. Moreover, in entirely different medical and biological datasets, the results are promising. For instance, on the medical PC dataset, the method shows a 1.79% improvement over the baseline framework, and a 2.07% improvement on the biological CG dataset. Hence, the experimental comparison results indicate that the proposed method demonstrates significant advantages in multiple dimensions. Additionally, the method also shows remarkable advantages in terms of resource consumption and training time.

To further validate the effectiveness and robustness of the proposed method, few-shot experiments were conducted on specific CrossNER, Res and Mov datasets in low-resource environments. The experimental results are presented in Tables 2 and 3.

The experimental results demonstrate that the proposed model consistently outperforms the baseline models in a small sample environment, with significant improvements observed under settings of K = 20 and K = 50. The method presented in this paper not only effectively leverages the rich semantic knowledge of LLMs but also addresses the issue of spurious correlations in LLMs through causal semantic relationships. This approach yields substantial improvements not only on the CrossNER dataset but also in scenarios where the source and target domains are entirely different. Specifically, compared to the best experimental results, the model achieves an average increase in F1-score by 1.73% and 1.50% under K = 20 and K = 50 sample settings, respectively. Furthermore, in the source and target domains (Res and Mov), where there are significant differences in domain and labels, the F1-scores for Res and Mov increase by 2.92% and 4. 83% at K = 20, and by 2.78% and 2.13% at K = 50, respectively. These results clearly indicate that the model effectively utilizes semantic features derived from causal relationships to enhance the rich semantic knowledge of LLMs.

By integrating features from both the source and target domains and identifying causal relationships between features, the model effectively mitigates the issue of spurious correlations between entities and context. This approach enhances the accuracy and reliability of the model in handling complex semantic tasks.

6 Experimental Analysis

In this subsection, the method selected the $ConLL03 \rightarrow CrossNER$ experimental setup, with the LLM used being LLaMA3.1-8B, to conduct a detailed analysis of the results from three perspectives: ablation experiments, parameter analysis, and case studies.

6.1 Ablation Analysis

To validate the effectiveness of each module, ablation experiments were conducted on the $ConLL03 \rightarrow CrossNER$ setup. The results are presented in Table 4. Where, $-L_y$ represents the removal of causal module information including E, C and R. It can be observed that the average F1-score decreases by 2.10% across different domains, with the highest decrease of 3.84% in the Sci domain. This indicates that the causal relationships between different entities and the physical world are crucial for the overall model performance. Removing this information diminishes the model's ability to verify and distinguish between different entity types and relationships. Consequently, this reduction negatively impacts the model's comprehensive performance in domain-specific tasks. Similarly, $-L_r$ represents the removal of feature module information. It is evident that the F1-score decreases by 1.57% on average across different domains, with the AI domain experiencing the highest reduction of 1.85%. This suggests that the integration of feature information provides effective data sources and real-world factual information, enhancing the model's adaptability and generalization capabilities across various domains. It also contributes to the overall performance of the model in cross-domain tasks.

Table 4: Ablation experiment (%)

Domain	Pol	Sci	Mus	Lit	AI
CDLM	74.12	73.41	79.32	71.23	64.44
$-L_y$	72.31	69.57	77.85	70.13	62.17
$-L_r$	73.48	72.12	78.33	70.87	62.59

Note: This table shows the results of the ablation experiment across different domains. CDLM represents the full model, while $-L_y$ and $-L_r$ represent ablations.

6.2 Parameter Analysis

To investigate the impact of the parameters λ_1 and λ_2 of the causal learning module on the experiment, the method set different parameters and conducted multiple trials. In the causal intervention and counterfactual modules, the loss function Loss is composed of L_e and L_{crd}, where λ_1 represents the weight of factual information, and λ_2 represents the weight of prior knowledge and complex causal relationships. The values for λ_1 and λ_2 were set to [0.2,0.4,0.6,0.8,1]. As shown in Fig. 4, when λ_2 is fixed, the increase in the value of λ_1 increases with the proportion of factual information, leading to an increase in the corresponding weight within the entire feature representation. At this point, it can be seen that as the weight of factual information increases, the model's ability to discriminate specific tasks improves accordingly, indicating that factual information can effectively enhance the model's accuracy and generalization capability. However, λ_1 is excessively large, the model may become overly reliant on factual information, neglecting prior knowledge

and complex causal relationships, leading to a decrease in overall performance. Similarly, if λ_1 is too small, the model cannot effectively utilize factual information, resulting in insufficient performance. Therefore, $\lambda_1 = 0.6$ and $\lambda_2 = 0.6$ as the optimal parameters for this set of experiments.

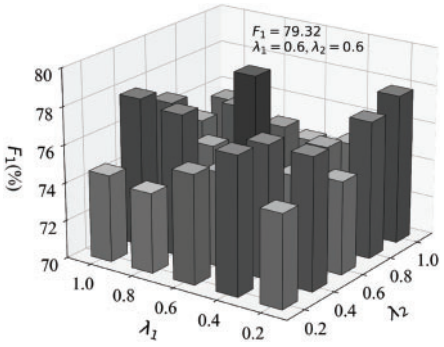

Figure 4: The values corresponding to different λ_1 and λ_2

6.3 Case Analysis

This study presents an in-depth case analysis by selecting representative sentences from various fields within the CrossNER dataset to elucidate the challenges and countermeasures in cross-domain tasks. The uniqueness of cross-domain tasks lies in the uncertainty and diversity of terminologies and labels across different domains. For instance, as demonstrated in Table 5, the term "Experts" might be labeled as "Person" in certain fields, while in others, it could be labeled as "Scientist". Such label uncertainties exacerbate the complexity of model processing and contribute to significant domain bias issues.

Table 5: Prediction comparison of different models in different fields

Input sentence	LightNER	CP-NER	CDLM
Ludwig van Beethoven was performing at the grand concert hall in Vienna.	Ludwig van Beethoven (Person) was performing at the grand concert hall in Vienna.	Ludwig van Beethoven (Person) was performing at the grand concert hall in Vienna.	Ludwig van Beethoven (Musician) was performing at the grand concert hall in Vienna.
In 1991, John Preskill and Kip Thorne bet against Stephen Hawking...	In 1991, John Preskill (Person) and Kip Thorne (Person) bet against Stephen Hawking (Person)...	In 1991, John Preskill (Person) and Kip Thorne (Person) bet against Stephen Hawking (Person)...	In 1991, John Preskill (Scientist) and Kip Thorne (Scientist) bet against Stephen Hawking (Scientist)...

In order to further analyze the predictive effectiveness of the model for different entity types, we performed a fine-grained analysis of the PC dataset, in which the main entity types in the PC dataset are Simple-chemical (CHEM), Gene-or-gene-product (GGP), Cellular-component (CCP) and Complex, respectively.

Table 6 shows the main experimental results, and it can be seen that compared with the LLM-based CP-NER method, CDLM improves its performance on several entity types, which once again validates the effectiveness of this paper's method. In addition, CDLM still achieves good performance even on CCP entity

types with less data volume, which further proves that CDLM has greater improvement even when facing the long-tail problem of data distribution.

Table 6: Fine-grained analysis (%)

Entity	CP-NER	CDLM
CCP	88.22	88.64
GGP	87.82	88.76
CHEM	88.86	89.62
Complex	88.02	89.58

Existing models often struggle with effectively identifying and distinguishing information from various unknown domains in cross-domain tasks. Although these models can accurately identify entity information, they exhibit a high degree of domain specificity in determining entity types, making generalization to unseen domains challenging. The method proposed in this paper leverages the extensive corpus of LLMs to integrate information from both source and target domains by exploiting cross-domain invariances. By incorporating causal learning, the model's capability to adapt to new domain entity information is further enhanced, resulting in more accurate predictions and entity type classifications. The model undergoes training and validation on multi-domain data, showcasing its robustness and effectiveness in CD-NER.

7 Large Model Analysis

7.1 Experimental Analysis

To evaluate the generalization capability of the LLMs, commonly used few-shot NER datasets were employed for testing, including ConLL-2003 and FewNERD. In Fig. 5, the main experimental results are summarized. Due to the high costs associated with using the ChatGPT API, Ma et al. [39]'s research was referenced for a comparative analysis of ChatGPT and InstructGPT results. The experimental findings indicate that in few-shot scenarios, the LLM performs exceptionally well in both 1-shot and 5-shot settings, significantly outperforming methods based on pre-trained language models. This superior performance is attributed to the LLM's extensive corpus and powerful generalization capabilities. However, as the number of samples increases, the performance curve gradually declines. This demonstrates the LLM's high sensitivity to small sample sizes and complex label information, indicating a challenge in fully leveraging limited training data to achieve accurate NER.

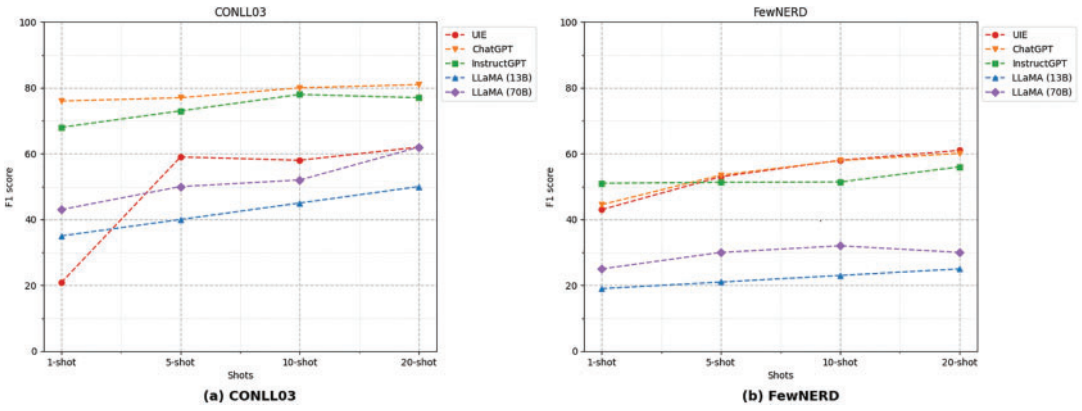

Figure 5: LLM experiment

7.2 Why LLM's Performance in Sequence Annotation Tasks Is Not Satisfactory

Through the above experiments, the performance deficiencies of LLM in sequence labeling tasks were analyzed, focusing on the following three key aspects: 1. Insufficient Utilization of Annotations: Compared to domain-specific models, LLMs gain limited benefits from increased training samples and label types, manifesting in two main constraints: (1). Effective Sample Capacity: The effective sample capacity is restricted by the model's maximum input length, leading to performance saturation before reaching the sample capacity limit. (2). Label Type Increase: An increase in label types results in fewer examples per label, limiting the LLM's ability to understand complex label interactions. This reduces the efficacy of annotations, as the LLM cannot leverage the expanded training data as effectively as domain-specific models. 2. Unfamiliarity with Task Format: LLMs are not sufficiently familiar with the highly flexible task formats typical in sequence labeling tasks, which impacts their contextual learning ability. The complex format of sequence labeling tasks, combined with a lack of relevant tasks in instruction-tuning datasets, makes it difficult for LLMs to accurately understand and execute these tasks. This unfamiliarity hampers the LLM's ability to handle diverse and intricate task structures. 3. Inadequate Contextual Understanding: LLMs exhibit limitations in handling complex contextual relationships and long-distance dependencies, affecting their performance in sequence labeling tasks. The model struggles to fully comprehend subtle contextual cues within the text, leading to insufficient labeling accuracy. These limitations hinder the LLM's ability to parse and understand nuanced information, which is crucial for precise sequence labeling.

8 Conclusion

In this paper, we propose leveraging causal relationships to enhance the CD-NER capability of LLMs. This approach mitigates domain-specific biases and enhances the model's adaptability to diverse tasks. Specifically, cross-domain feature fusion representations are improved by effectively incorporating causal relationships between entities and their contexts. Experimental results validate the model's effectiveness, demonstrating superior performance across various domains and providing novel insights for future research. Despite its robust performance, the model still has some limitations. Future work will focus on enhancing rapid adaptation to new domains by optimizing cross-domain feature integration, and optimizing model algorithms by analyzing information such as the computational complexity of the model. There is also significant value in introducing simpler baseline methods that provide models with more flexible adaptability and higher computational efficiency. At the same time, exploring how to effectively fine-tune the causality of techniques using parameters is critical to better capture and utilize domain-specific features, which will further enhance the ability of models to handle a variety of complex tasks.

Acknowledgement: None.

Funding Statement: This research was supported by National Natural Science Foundation of China Joint Fund for Enterprise Innovation Development (U23B2029), National Natural Science Foundation of China (62076167, 61772020), Key Scientific Research Project of Higher Education Institutions in Henan Province (24A520058, 24A520060, 23A520022) and Postgraduate Education Reform and Quality Improvement Project of Henan Province (YJS2024AL053).

Author Contributions: The authors confirm contribution to the paper as follows: study conception and design: Jiahao Wu, Jinzhong Xu, Xiaoming Liu, Jie Liu; data collection: Jiahao Wu, Guan Yang; analysis and interpretation of results: Jiahao Wu, Jinzhong Xu; draft manuscript preparation: Jiahao Wu. All authors reviewed the results and approved the final version of the manuscript.

Availability of Data and Materials: The data that support the findings of this study are openly available in CrossNER at https://github.com/zliucr/CrossNER (accessed on 16 February 2025).

Ethics Approval: Not applicable.

Conflicts of Interest: The authors declare no conflicts of interest to report regarding the present study.

Nomenclature

LLMs	Large Language Models
CD-NER	Cross-Domain Named Entity Recognition
NER	Named Entity Recognition
NLP	Natural Language Processing
EB	Entity Boundaries
LB	Label Boundaries
IE	Information Extraction
GCN	Graph Convolutional Network
CD-LM	Cross-Domain Language Model
BCE	Binary Cross Entropy

References

1. Brown TB, Mann B, Ryder N. Language models are few-shot learners. arXiv:2005.14165. 2020.
2. Alqaaidi SK, Bozorgi E. A survey on recent named entity recognition and relation classification methods with focus on few-shot learning approaches. arXiv:2310.19055. 2023.
3. Wei J, Wang X. Chain of thought prompting elicits reasoning in large language models. arXiv:2201.11903. 2022.
4. Chen Z, Xu L, Zheng H, Chen L, Tolba A, Zhao L, et al. Evolution and prospects of foundation models: from large language models to large multimodal models. Comput Mater Contin. 2024;80(2):1753–808. doi:10.32604/cmc.2024.052618.
5. Kuang K. Causal inspired trustworthy machine learning. In: Proceedings of the ACM Turing Award Celebration Conference—China 2023. ACM TURC '23; 2023; New York, NY, USA: Association for Computing Machinery. p. 3–4. doi:10.1145/3603165.3607365.
6. Tao Z, Jin Z, Bai X. SEAG: structure-aware event causality generation. In: Findings of the Association for Computational Linguistics: ACL 2023; 2023; Toronto, ON, Canada: Association for Computational Linguistics. p. 4631–44.
7. Lin J, Zhou J, Chen Q. Causal intervention-based prompt debiasing for event argument extraction. arXiv:2210.01561. 2022.
8. Ren L, Liu Y, Cao Y, Ouyang C. CoVariance-based causal debiasing for entity and relation extraction. In: Bouamor H, Pino J, Bali K, editors. Findings of the association for computational linguistics: EMNLP 2023. Singapore: Association for Computational Linguistics; 2023. p. 2627–40.
9. Hu J, Zhao H, Guo D. A label-aware autoregressive framework for cross-domain NER. In: Findings of the Association for Computational Linguistics: NAACL 2022; 2022; Seattle, WA, USA: Association for Computational Linguistics. p. 2222–32.
10. Yang Z, Liu Y, Ouyang C. Causal intervention-based few-shot named entity recognition. In: Findings of the Association for Computational Linguistics: EMNLP 2023; 2023; Singapore: Association for Computational Linguistics. p. 15635–46.
11. Kalyan KS. A survey of GPT-3 family large language models including ChatGPT and GPT-4. arXiv:2310.12321. 2023.
12. Mou L, Meng Z, Yan R. How transferable are neural networks in NLP applications? In: Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing; 2016; Austin, TX, USA: Association for Computational Linguistics. p. 479–89.

13. Tang M, Zhang P. DoSEA: a domain-specific entity-aware framework for cross-domain named entity recogition. In: Proceedings of the 29th International Conference on Computational Linguistics; 2022; Gyeongju, Republic of Korea: International Committee on Computational Linguistics. p. 2147–56.
14. Li X, Feng J. A unified MRC framework for named entity recognition. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics; 2020; Online, Association for Computational Linguistics. p. 5849–59.
15. Wang J, Kulkarni M. Multi-domain named entity recognition with genre-aware and agnostic inference. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics; 2020; Association for Computational Linguistics. p. 8476–88.
16. Das SSS, Katiyar A. CONTaiNER: few-shot named entity recognition via contrastive learning. In: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers); 2022; Dublin, Ireland: Association for Computational Linguistics. p. 6338–53.
17. Xu J, Zheng C, Cai Y. Improving named entity recognition via bridge-based domain adaptation. In: Findings of the Association for Computational Linguistics: ACL 2023; 2023; Toronto, ON, Canada: Association for Computational Linguistics. p. 3869–82.
18. Han R, Yang C, Peng T. An empirical study on information extraction using large language models. arXiv:2305.14450. 2024.
19. Qin C, Zhang A, Zhang Z. Is ChatGPT a general-purpose natural language processing task solver? arXiv:2302.06476. 2023.
20. Fei Y, Hou Y, Chen Z. Mitigating label biases for in-context learning. In: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers); 2023; Toronto, ON, Canada: Association for Computational Linguistics. p. 14014–31.
21. Wang F, Mo W, Wang Y. A causal view of entity bias in (Large) language models. In: Findings of the Association for Computational Linguistics: EMNLP 2023; 2023; Singapore: Association for Computational Linguistics. p. 15173–84.
22. Ye J, Xu N, Wang Y, Zhou J, Zhang Q, Gui T, et al. LLM-DA: data augmentation via large language models for few-shot named entity recognition. arXiv:2402.14568. 2024.
23. Jimenez Gutierrez B, McNeal N, Washington C. Thinking about GPT-3 In-context learning for biomedical IE? Think again. In: Findings of the Association for Computational Linguistics: EMNLP 2022; 2022; United Arab Emirates: Abu Dhabi. p. 4497–512.
24. Zhang K, Jimenez Gutierrez B, Su Y. Aligning instruction tasks unlocks large language models as zero-shot relation extractors. In: Findings of the Association for Computational Linguistics: ACL 2023; 2023; Toronto, Canada: Association for Computational Linguistics. p. 794–812.
25. Tang K, Niu Y, Huang J. Unbiased scene graph generation from biased training. arXiv:2002.11949. 2020.
26. Lin Z, Ding H, Hoang NT. Pre-trained recommender systems: a causal debiasing perspective. arXiv:2310.19251. 2024.
27. Cao B, Lin H, Han X. Can prompt probe pretrained language models? Understanding the invisible risks from a causal view. In: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers); 2022; Dublin, Ireland: Association for Computational Linguistics. p. 5796–808.
28. Zheng J, Chen H, Ma Q. Cross-domain Named entity recognition via graph matching. In: Findings of the Association for Computational Linguistics: ACL 2022; 2022; Dublin, Ireland: Association for Computational Linguistics. p. 2670–80.
29. Tjong Kim Sang EF, De Meulder F. Introduction to the CoNLL-2003 Shared Task: language-Independent Named Entity Recognition. In: Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003; 2003. p. 142–7.
30. Bossy R, Kim JD, Kim JJ. Overview of BioNLP shared task 2013. In: Proceedings of the BioNLP Shared Task 2013 Workshop; 2013; Sofia, Bulgaria: Association for Computational Linguistics. p. 1–7.
31. Liu J, Pasupat P, Cyphers S, Glass J. Asgard: a portable architecture for multilingual dialogue systems. In: 2013 IEEE International Conference on Acoustics, Speech and Signal Processing; 2013; Vancouver, BC, Canada. p. 8386–90.

32. Liu Z, Xu Y, Yu T, Dai W, Ji Z, Cahyawijaya S, et al. CrossNER: evaluating cross-domain named entity recognition. arXiv:2012.04373. 2020.
33. Devlin J, Chang MW, Lee K. BERT: pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers); 2019; Minneapolis, MN, USA: Association for Computational Linguistics. p. 4171–86.
34. Liu Z, Winata GI. Coach: a coarse-to-fine approach for cross-domain slot filling. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics; 2020; Association for Computational Linguistics. p. 19–25.
35. Yang Y, Katiyar A. Simple and effective few-shot named entity recognition with structured nearest neighbor learning. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP); 2020; Association for Computational Linguistics. p. 6365–75.
36. Chen X, Li L, Deng S. LightNER: a lightweight tuning paradigm for low-resource NER via pluggable prompting. In: Proceedings of the 29th International Conference on Computational Linguistics; 2022; Gyeongju, Republic of Korea: International Committee on Computational Linguistics. p. 2374–87.
37. Ma R, Zhou X, Gui T, Tan Y. Template-free prompt tuning for few-shot NER. In: Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies; 2022; Seattle, WA, USA: Association for Computational Linguistics. p. 5721–32.
38. Chen X, Li L, Qiao S. One model for all domains: collaborative domain-prefix tuning for cross-domain NER. arXiv:2301.10410. 2023.
39. Ma Y, Cao Y, Hong Y. Large language model is not a good few-shot information extractor, but a good reranker for hard samples! In: Findings of the Association for Computational Linguistics: EMNLP 2023; 2023; Singapore: Association for Computational Linguistics. p. 10572–601.